2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Ranking news articles based on popularity prediction

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida
LIP6/CNRS – UPMC Sorbonne Universités
4 Place Jussieu, 75005, Paris, France
Tel/Fax: +33 (0) 144278877 / +33 (0) 14425353
Emails: {tatar, antoniad, amorim, sf}@npa.lip6.fr

***Abstract*—News articles are a captivating type of online content that capture a significant amount of Internet users' interest. They are particularly consumed by mobile users and extremely diffused through online social platforms. As a result, there is an increased interest in promptly identifying the articles that will receive a significant amount of user attention. This task falls under the broad scope of content popularity prediction and has direct implications in various contexts such as caching strategies or online advertisement policies. In this paper we address the problem of predicting the popularity of news articles based on user comments. We formulate the prediction task into a ranking problem where the goal is not to infer the precise attention that a content will receive but to accurately rank articles based on their predicted popularity. To this end, we analyze the ranking performance of three prediction models using a dataset of articles covering a four-year period and published by 20minutes.fr, an important French online news platform. Our results indicate that prediction methods improve the ranking performance and we observed that for our dataset a simple linear prediction method outperforms more dedicated prediction methods.**

***Index Terms*—News articles, ranking, prediction, popularity**

I. Introduction

News articles are a type of content that can be easily produced, have a small size, short lifespan, and low cost, properties that makes them interesting for fast information diffusion through social media platforms or social networks. As a consequence, a significant amount of research has been focused in understanding the interest around news, including general observation on how content is generated, describing the decay of interest over time, community detection, and prediction of popularity. It is the former one, however, that gained most of the research focus both because this problem is very challenging and for its immediate practical implications, where predicting the popularity of online content is valuable for different actors: news sites and news aggregators can better highlight their popular content, online advertisers could propose more profitable monetization strategies, and online readers can filter more easily the huge amount of information.

There are different ways of expressing the notion of popularity. For example, the classical way of defining it is through the number of views. However, this information is seldom available by external observers and when available it is difficult to estimate the actual number of times that the page was requested by users or due to web crawlers and search engines. Nevertheless, as reading news has become a social experience, there are other metrics that indicate if a content is popular or not. These metrics are based on user participation activity such as user comments, votes, or sharing through social media or email services. In this paper we focus on one dimension of the content popularity and consider the *number of comments* as an implicit evaluator of the interest generated by an article.

Predicting the popularity of news articles is a complex and difficult task and different prediction methods and strategies have been proposed in several recent studies [1], [2], [3], [4]. The common point of all these methods is that they focus on predicting the *exact* attention that an article will generate; thus, one prediction method is preferred over another depending on the type of error that we want to minimize [1]. Indeed, this is useful for online advertising where the revenues would be calculated per exact attention that an article will receive. However, in other practical situations, it is not the exact amount of attention that we are interested in, but *how popular an article is relatively to the others*. This can be more formally defined as the problem of *ranking online news*, an ability of the prediction methods that has not been previously studied.

In this paper we investigate the feasibility of using prediction methods to rank news articles based on their predicted popularity. To this end, we compare the ranking capabilities of three prediction methods: a linear model, a linear model on a logarithmic scale, and a constant scaling model. In order to properly evaluate the relevance of the ranking strategy, we propose a general setting that takes into consideration two important properties of the articles: lifetime and distribution of popularity. We have performed our analysis on a complete and significantly large dataset, covering a four-year period, that contains all the articles and comments published on 20minutes.fr.

II. Dataset

A. Dataset source

The dataset that we have used in this paper consists of all articles and comments published from February 2007 to January 2011 on 20minutes.fr, the third most visited news website in France[1]. A data cleaning has been performed to remove articles for which the commenting feature has been

[1] http://www.mediametrie.fr

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.28

disabled in the end relying for our analysis on 260,000 articles and 2,600,000 comments. In the following we would like to briefly highlight two important characteristics of this dataset that are relevant to the prediction evaluation: the lifetime of the articles and the distribution of popularity. The lifetime gives us an intuition of how user attention in news articles fades over time and allows us to do a more rigorous ranking evaluation. In the same way, understanding how attention is distributed over the articles will permit us to do a more focused evaluation, where a prediction method should be particularly accurate in identifying the top most important articles.

B. Lifetime of an article

A common characteristic of online content is that it suffers from a decay of interest over time, and depending on the type of content, this interest may be steep or gradual. News articles depict a very steep decay as they refer to a recent type of information that by its nature has a limited lifespan of interest. We provide a coarse evaluation of how the interest in 20minutes articles decreases over time by analyzing the probability of an article to receive comments a certain time after its publication.[2]

We represent this through a complementary cumulative distribution function of the last comment received by articles (Figure 1). As we can observe, there is a 40% probability that an article will receive comments one day after its publication and this probability drops up to 20% after three days. A possible explanation comes from the fact that from all comments posted on daily basis, a significant amount of 90% aim the articles published at most 24 hours before and only 2% go to articles published earlier than three days.

There are, however, articles that continue to receive attention after this period, but in most of the cases they represent only a sparse interest and not a constant one as observed for other type of traffic [7]. This latent interest can be explained by users' use of search engines or as a result of web links posted on forums or online social networking sites.

These results suggest that a one day observation period is a good indication of articles' popularity and therefore the prediction methods should perform particularly well for articles not older than one day.

C. Distribution of popularity

Similar to most of the content found on the Internet, it has been observed that the popularity of online articles follows a highly skewed distribution of popularity [2], [5], [8]. This observation is also a characteristic of our data. Using the statistical tests proposed in [9], we have observed that our data depicts a power-law characteristics in the tail of the distribution[3]. This can visually be observed in Figure 2, where we present the complementary cumulative distribution

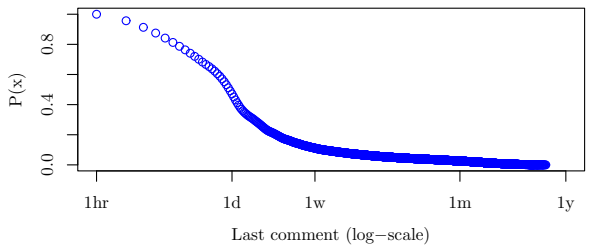

Fig. 1. Complementary cumulative distribution function corresponding to the articles' lifetime (time elapsed between article publication time and the last comment time). The x-axis tics correspond to one **h**ou**r**, **d**ay, **w**eek, **m**onth, and **y**ear.

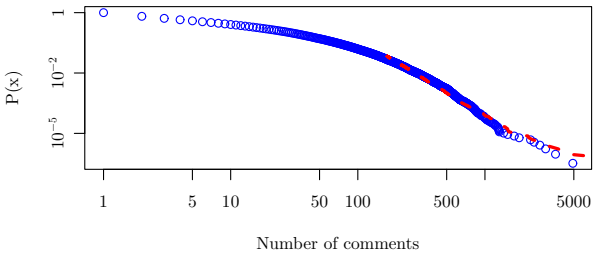

Fig. 2. The complementary cumulative distribution function of the articles' popularity and its corresponding power-law fit.

function of the number of comments received per article and the corresponding power-law fit. Regarding the distribution of popularity on a daily basis, our results indicate that 20% of the articles receive 80% of all the comments posted that day, where all these articles have been published at most one day before.

The power-law observation can be explained by the recommendation strategy proposed by 20minutes. The website highlights the most commented articles in a dedicated section and twice a day it sends to its subscribers a short edition with the most commented articles. This creates a *rich-get-richer* effect, which is one of the reasons why power-law appears so often on the Internet [10]. The recommendation mechanism can also explain why the power-law fails to appear in the beginning of the distribution. Articles that are unpopular in the beginning do not benefit from any recommendation mechanism and the probability of receiving any kind of attention drops even more as they loose their position on the website [6].

III. Prediction Models

In this paper we compare the ranking capabilities of three prediction models:

[2]We are aware that there are other fine-grained methods of evaluating the decay of attention over time [3], [5], [6], but for the scope of our work, this coarse characterization provides us with sufficient information.

[3]Statistical techniques based on maximum-likelihood methods and Kolmogorov-Smirnov statistics.

- A simple **linear regression** model that we have proposed in our preliminary analysis of the predictive characteristics on 20minutes articles [11].
- The **linear regression on a logarithmic scale** model proposed by Szabo and Huberman [1] and previously evaluated on Digg news, Youtube videos, and news articles from seven Dutch online news websites [2].
- The **constant scaling** model described by Szabo and Huberman [1] and evaluated on Digg news and Youtube videos.

The first model has been considered for its simplicity and can be seen as a baseline model used for estimating the improvement that one can obtain when using more specialized prediction models. The other two prediction models are more specialized functions that have already shown good results in predicting the popularity of different online content. The linear model on a logarithmic scale is a model that is well adapted to data that presents heavy tail characteristics. We have also considered the constant scaling model in our analysis following the observations that this model outperforms the previous one if the prediction goal is to minimize the relative squared error [1].

The three models listed above are all regression functions in which the dependent variable is the total number of comments that an article s will receive at a later time t_r and the independent variable is the number of comments received t_i hours after its publication ($t_i < t_r$). The goal of the prediction methods is then to estimate the number of comments received t_r hours after the publication of an article using the number of comments received in the first t_i hours. We set t_r to 30 days[4] and vary t_i from 1 to 24 hours.

In the case of the *simple linear model* the coefficients of the function, α_1 and β_1 , are obtained through the method of least squares and the function has the following form:

$$\widehat{N}_s^{\mathrm{LM}}(t_i, t_r) = \alpha_1(t_i, t_r) + \beta_1(t_i, t_r) \times N_s(t_i)\,. \qquad (1)$$

The second model, the *linear model on a logarithmic scale* is expressed by the following equation:

$$\widehat{N}_s^{\mathrm{LN}}(t_i, t_r) = \exp(\ln(N_s(t_i)) + \beta_0(t_i, t_r) + \sigma_0^2(t_i, t_r)/2). \quad (2)$$

The coefficient of this equation, β_0, is obtained through maximum likelihood estimation and σ_0^2 is the estimate of the variance of the residuals on a logarithmic scale.

The last model, called *constant scaling*, is expressed as

$$\widehat{N}_s^{\mathrm{CS}}(t_i, t_r) = \alpha_2(t_i, t_r) \times N_s(t_i)\,, \qquad (3)$$

where we estimate α_2 using the following expression:

$$\alpha_2(t_i, t_r) = \frac{\sum_s \frac{N_s(t_i)}{N_s(t_r)}}{\sum_s \left[\frac{N_s(t_i)}{N_s(t_r)}\right]^2}\,. \qquad (4)$$

[4]We have observed that 98% of articles receive all their comments within this time period.

IV. Ranking performance

A. Evaluation methodology

For our evaluation we separate our data into a training and a test set. We obtain the prediction models on the training set, using the articles published between 2007 and 2009, and evaluate the ranking accuracy on the test set that contains all articles published in 2010. We evaluate the ranking performance of the prediction methods using the following strategy. We set a prediction hour t_p, time at which we predict the popularity of all articles published in the last 24 hours and rank them based on the predicted value. We will further refer to this ordering as the *predicted ranking* and define the ground truth ranking as the ordering based on the real number of comments that the articles received t_r hours after their publication. We use two error metrics to compare the two rankings: Kendall rank correlation and Mean Average Precision (MAP). The first evaluation metric, Kendall rank correlation coefficient, is a non-parametric statistical test that measures the similarity between two independent orderings, in our case between the true and predicted ranking. It ranges from -1 to 1, where 1 means there is a perfect agreement between the two rankings. We use Kendall to compare the full agreement between the two rankings but as we have seen in Section II just a small percent of articles gather most of the attention and in this case we are interested in the performance of the prediction methods for the top-k articles, for small values of k. We evaluate this ranking performance using a common measure used in the field of information retrieval, the Mean Average Precision (MAP)[12].

The average precision (AP) at level k for one ranking is defined as:

$$AP_k = \frac{\sum_{1 \le i \le k} P(i) \times rel(i)}{k}\,, \qquad (5)$$

where *P(i)* is the precision at level *i*, and *rel(i)* $= 1$ if the article found at rank *i* is relevant and 0 otherwise. MAP_k is the mean value for all AP_k over the entire test set.

We evaluate the three prediction models and compare them with four other heuristics:

- *Random* model: we randomly choose the articles that will be the most popular in the future.
- *Current* popularity model: the articles that were the most popular at the prediction moment t_p will also be the most popular in the future.
- Time of publication model (*freshness*): the further the article has been published relative to the prediction moment t_p the less popular it will be in the future.
- A *weighted* model between the time of publication and the number of comments: the more number of comments received in a shorter period is be a good indication of the future popularity.

B. Results

In the following we present some relevant results of the ranking evaluation. In Figure 3, by means of a box plot, we depict the ranking performance in terms of Kendall rank

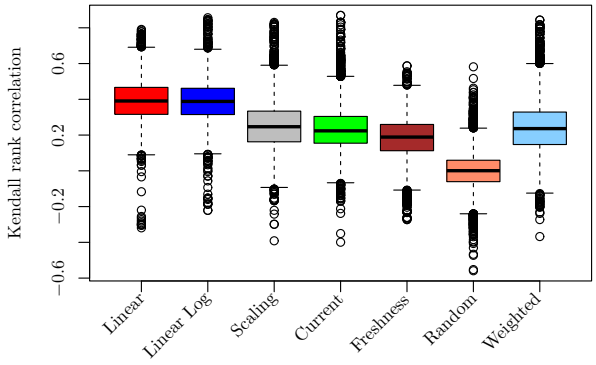

Fig. 3. Kendall rank correlation corresponding to a scenario where on a daily basis and for all hours of the day we evaluate the ranking accuracy of the three prediction models and simple heuristics.

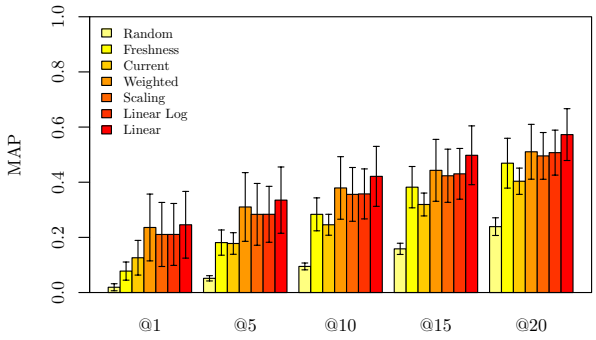

Fig. 4. MAP at different levels of precision. @n is the MAP score for the top most important n articles. We present the mean value and 95% confidence intervals for all prediction hours.

correlation. The box plot is a representation of all Kendall correlation coefficients obtained on the test set where, on a daily basis and for all hours of the day, we evaluate the ranking accuracy using the previously described strategy. We can observe from the figure that if ranking is based on simple heuristics there is a weak positive correlation between the true ranking and the predicted one and, surprisingly, there is no improvement of the ranking accuracy if we use the constant scaling prediction model. A fair improvement of the ranking can, however, be observed in the case of the two linear prediction functions, which show a moderate positive correlation between the true and the predicted ranking.

Given our previous observations of how the most popular articles obtain an impressive percent of daily user attention, we compare the performance of the models when faced with the challenge of identifying the top most important articles. In order to do this, we analyze the MAP score for 5 levels of precision: MAP@1, MAP@5, MAP@10, MAP@15, MAP@20. The results are presented in Figure 4 and correspond to the mean value with a 95% confidence interval of MAP scores for all prediction hours. The first observation that we make is that a simple linear prediction model outperforms the two other prediction models and simple heuristics for all levels of precision. In general, all prediction methods show better performances than the *current* popularity model or *freshness* model and what is particularly encouraging is that the prediction methods show a significant improvement in identifying the most important and the top five most important articles. It can also be observed from the figure that the *weighted* model can be a good alternative for this prediction task, as it generally outperforms the constant scaling and the linear log model.

We conclude our analysis with one final observation on the sensitivity of the ranking to the prediction hour. In our previous analysis ([11]), and other similar studies, we have observed that articles and comments are generated at a different rate during the day. As a consequence, articles may be more popular or exhaust their interest more quickly depending on the publication hour, which may influence the ranking accuracy. We explore the possible dependency between the ranking performance and the circadian pattern of content generation by plotting the ranking accuracy for different prediction hours. We present the results in Figure 5, where we plot the ranking accuracy of the prediction methods for MAP@20 on an hourly basis. The figure indicates that the ranking accuracy is insignificantly influenced by the prediction hour, where the relative performance of each one of the prediction methods holds on an hourly basis with the linear model showing the highest accuracy in correctly identifying the top most important articles.

V. Related Work

Several works have addressed the problem of predicting the popularity of online content. Kaltenbrunner et al. [13] have proposed a model that considers a uniform growth of the popularity, depending on the publication hour, and used it in the context of predicting the number of comments for Slashdot stories. Szabo et al. [1] have proposed two other prediction methods that have shown good results in predicting the popularity of Youtube videos and Digg stories. Tsagkias et al. [2] showed that the linear log method proposed in [1] is an adequate method for predicting the popularity of news articles. Lerman et al. [4] present a different prediction model using data in the form of social influence and web platform characteristics. A different approach has been proposed by Lee et al. [3] where, instead of predicting the exact value, the authors are interested in predicting the probability that a content will continue to receive comments after a certain period of time. More recent results [14], which use the number of tweets as the popularity metric, show that it is possible to classify articles in four classes of popularity but still difficult

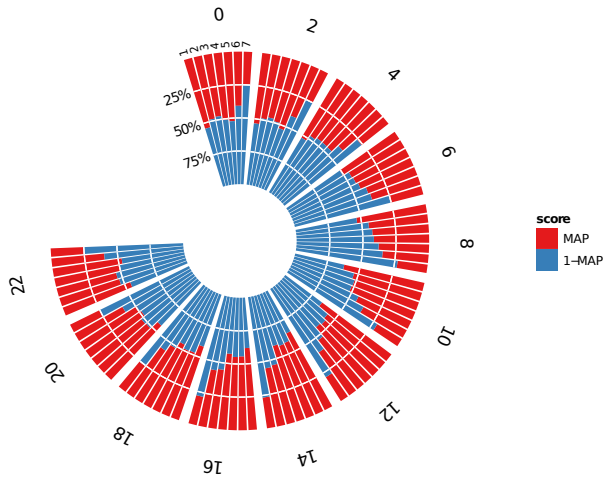

Fig. 5. The performance of each model in predicting the most important 20 articles (MAP@20). The outer numbers correspond to different reference hours t_p. The inner numbers correspond to the prediction methods and heuristics where 1 - linear, 2 - linear log, 3 - constant scaling, 4 - weighted model, 5 - freshness, 6 - current number of comments, 7 - random.

to predict the exact amount of attention. We place ourselves in this context of popularity prediction. In our work we analyze the predictive characteristics of news articles, on an unexplored dataset, using methods that have showed good results in previous works. We make a step further in our research and analyze the ranking capabilities of these methods by taking into consideration the dynamic nature of news generation.

The feasibility of ranking online content has been addressed in [15], [16]. Hsu et al. [15] have studied the possibility of ranking comments inside news stories, based on the votes that they receive and using a machine learning technique. In a different context Yin et al. [16] proposed a ranking model, based on user characteristics and by combining positive and negative votes, that allows one to rank the potential popular funny stories inside JokeBox. In our study we share the same general objective of correctly ranking popular items, but our work differs both in the type of content that we are analyzing and the ranking technique.

VI. Conclusions and Future Work

In this paper, we analyzed the capacity of three prediction methods in predicting the exact ordering of articles based on their future popularity. We conducted our analysis on a complete data set of articles and comments from an important French online news platform. Throughout our evaluation we have considered two important properties of news articles, the distribution of popularity and the lifetime of articles. Our results indicate that for this particular prediction task the most appropriate prediction method, out of the three that we have analyzed and for this dataset, is a simple linear regression. From a general point of view we have observed that prediction methods do have an impact on the ranking accuracy, but their performance is rather limited giving that a simple heuristic, that includes the creation time of articles and the partial number of comments, reveals an accuracy that is comparable to the accuracy obtained when using prediction methods.

There are several directions that we will consider for our future work. Our first step would be to evaluate the effectiveness of dedicated learning to rank algorithms, that are extensively used in information retrieval evaluation and compare them with the prediction methods used in this paper. A second direction consists in an evaluation of the ranking accuracy for other datasets, which will give us insights on the extent to which one can generalize the prediction methods for a wide variety of online content. Finally, all these observations would be meaningless if not evaluated in a practical context. We will therefore consider evaluating the impact of these ranking methods on caching strategies used in realistic scenarios.

References

[1] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, p. 80, 2008.

[2] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ser. ECIR2010. Springer, 2010.

[3] J. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. IEEE Computer Society, 2010.

[4] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 621–630.

[5] F. Wu and B. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, p. 17599, 2007.

[6] M. Simkin and V. Roychowdhury, "Why does attention to web articles fall with time?" *Arxiv preprint arXiv:1202.3492*, 2012.

[7] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.

[8] P. Van Mieghem, N. Blenn, and C. Doerr, "Lognormal distribution in the digg online social network," *Eur. Phys. J. B*, vol. 83, pp. 251–261, 2011.

[9] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, November 2009.

[10] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[11] A. Tatar, P. Antoniadis, A. Limbourg, M. D. de Amorim, J. Leguay, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *WIMS'11*. ACM, 2011, pp. 67–75.

[12] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[13] A. Kaltenbrunner, V. Gomez, and V. Lopez, "Description and prediction of slashdot activity," in *Proceedings of the 2007 Latin American Web Conference*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 57–66.

[14] R. Bandari, S. Asur, and B. Huberman, "The pulse of news in social media: Forecasting popularity," *Arxiv preprint arXiv:1202.0332*, 2012.

[15] C. Hsu, E. Khabiri, and J. Caverlee, "Ranking comments on the social web," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4. IEEE, 2009, pp. 90–97.

[16] P. Yin, P. Luo, M. Wang, and W.-C. Lee, "A straw shows which way the wind blows: ranking potentially popular items from early votes," in *Proceedings of the fifth ACM international conference on Web search and data mining*, ser. WSDM '12. ACM, 2012, pp. 623–632.